Report 3

LI, Ruizhe | 1155076990

Github: <https://github.com/rzli6/ML-Storage.git> (private)

Target Paper: Predicting Disk Replacement towards Reliable Data Centers

Goal: Realize the method descripted in the paper.

Dataset: 2015 Q1 – 2015 Q4

Apply RGF model trained on ST4000DM000 (SgtA) on SgtA:

F1: 0.9948348265944833

Recall: 0.9897725626539187

Precision: 1.0

Apply RGF model to *ST31500541AS* **without** transfer learning:

F1: 0.12097669256381798

Recall: 1.0

Precision: 0.064

Note: It nearly predicts all the test cases to be failed, which is far from the truth.

Transfer Learning:

Knowing that the original f1 score is so low it might be meaningless to do transfer learning, I still tried the procedure described in the paper. I merged the SgtA and SgtB datasets together, and trained a LR model to classify the two models. There are 1693 disks in SgtB and 1586 disks in SgtA, totally 3279 disks in the dataset. The trained model can reach a high precision of 99%. Usually we expect a higher score, but under this situation, a high score means the two disks are so dissimilar to each other that the model can easily tell the difference. Therefore, I couldn’t go any further to select similar cases for SgtA model.

Apparently, I failed on transfer learning. And the conclusion is that, we cannot carry it out until we know how to get a higher mark when applying the original learner on the new model. Until then, we could train a classifier to help us select samples similar to SgtB.

Target Paper: Improving Storage System Reliability with Proactive Error Prediction

Goal: Realize the method descripted in the paper.

Dataset: 2015 Q1 – 2015 Q4

Dataset Overview:

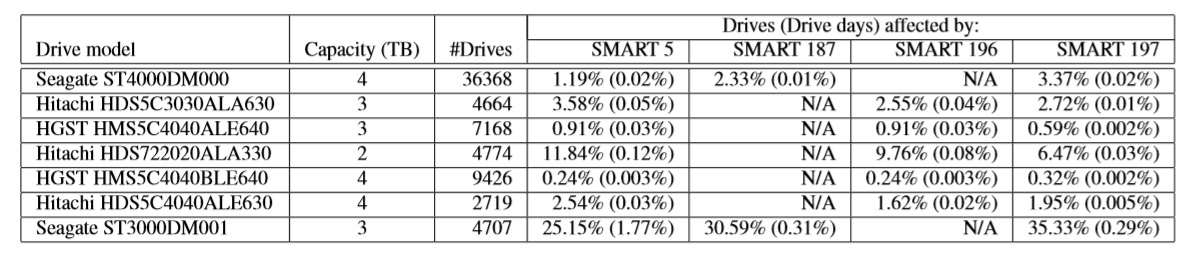
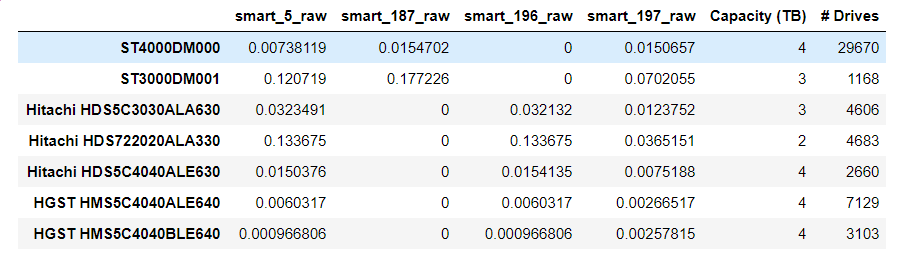


Figure in the paper



My result

Questions:

1. What we are going to predict? (Smart\_5\_raw or the increasement of it, or something else?)
2. Should we include the smart\_5\_raw as a feature? As shown in the paper?

Following is my result upon my understanding.

Label: 1 if smart\_5\_raw increases in the next week, else 0

Features (20 in total):

['y',

'smart\_1\_raw', 'smart\_4\_raw', 'smart\_5\_raw',

'smart\_7\_raw', 'smart\_9\_raw', 'smart\_12\_raw',

'smart\_187\_raw', 'smart\_193\_raw', 'smart\_194\_raw',

'smart\_197\_raw', 'smart\_199\_raw', 'smart\_4\_raw\_increase',

'smart\_5\_raw\_increase', 'smart\_7\_raw\_increase',

'smart\_9\_raw\_increase', 'smart\_12\_raw\_increase',

'smart\_187\_raw\_increase', 'smart\_193\_raw\_increase',

'smart\_197\_raw\_increase', 'smart\_199\_raw\_increase']

Note that, after data preprocessing I only got 135 cases with a label 1.

Down sampling to 500 health samples with KMeans.

Then feed the 635 training data into learner, I got:

